Ethics of Graphing

Principles of Graphing

Dr. Michael E. Olson

MATH 3080 - Foundations of Data Science

Methods of Graphing

- Variety of Graphs
  - Lineplot, Scatterplot, Bar graph, Boxplots, Histograms and KDE plots

- Tools
  - Matplotlib, Seaborn, Plotly

Mistake or Misleading?

Some graphs give readers incorrect impressions.

Sometimes, people honestly make mistakes

Sometimes, people purposefully use bad statistics to mislead readers

Example of Misleading Statistics

In the 2020 election season, a news article said:

> Older, white voters are significantly more likely to vote by mail and have those ballots counted, studies show, while voters of color and younger voters are significantly more likely to have their ballots rejected.
> NBC News, Aug 9, 2020

Could this be deceptive?

Example of Misleadding Statistics

It likely isn't purposefully deceptive - just a wording issue. But here is a scenario where their statement is true, but completely misleading.

	White	Non-White
Older	45%	25%
Younger	20%	10%

Younger AND non-white groups would be 20%+25%+10%=55%, a majority, though the biggest problem is older white voters.

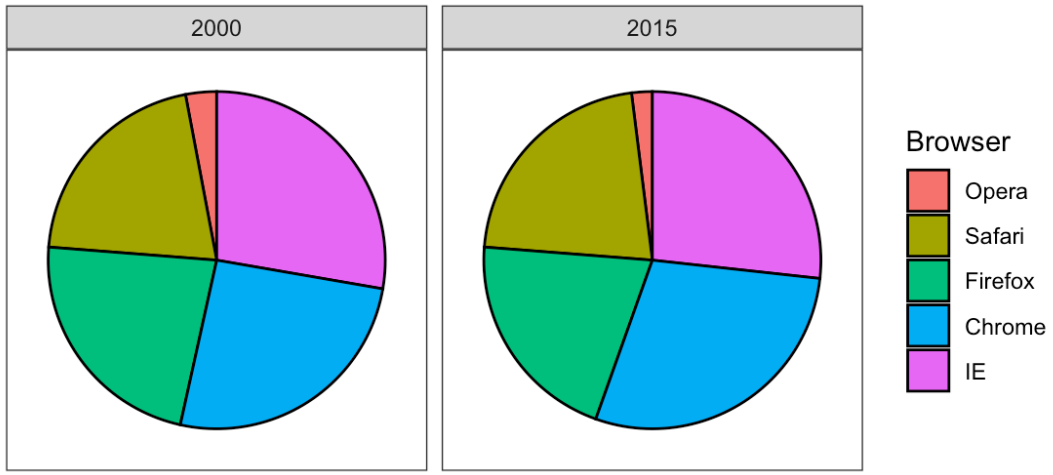
Mistake or Misleading?

In the following slides, we will look at a few principles that could convey information incorrectly and how to avoid them.

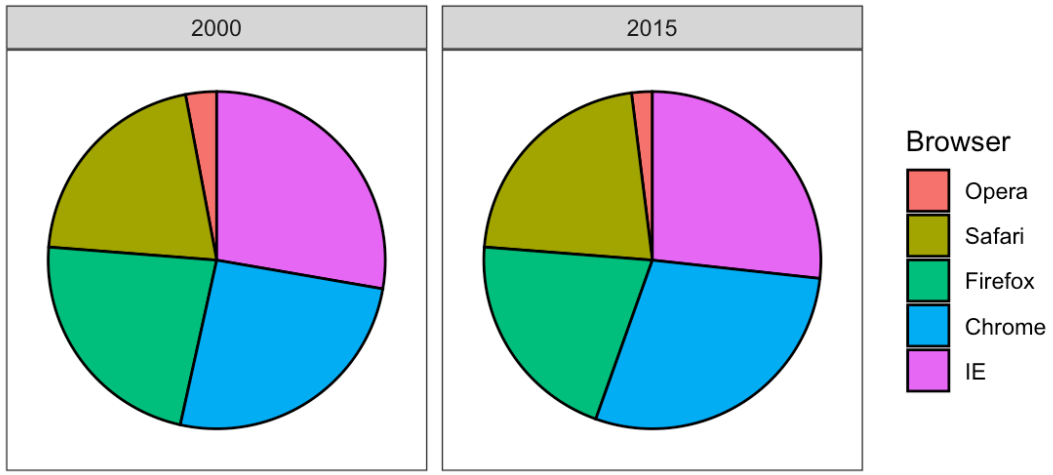
Note that these figures come from our textbook:

- Irizarry, *Introduction to Data Science*, Chapter 9

Visual Cues

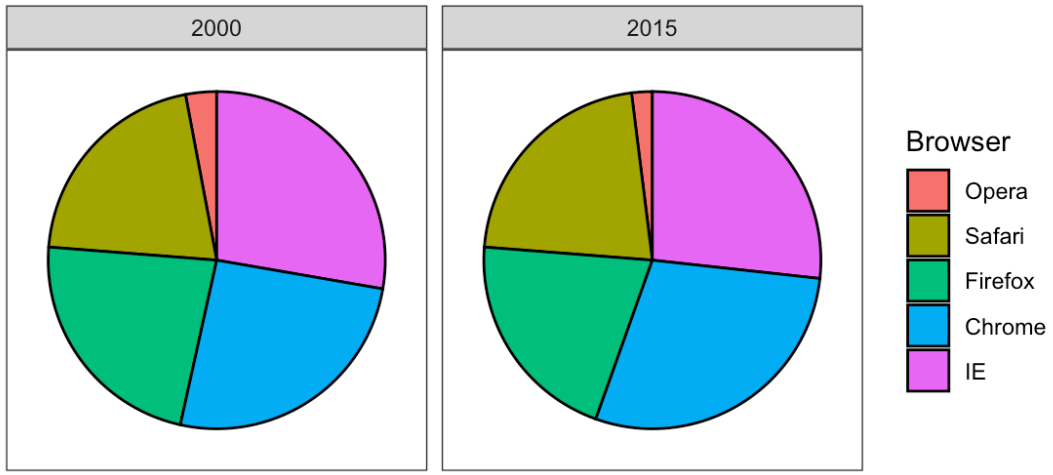


Visual Cues



- Which had the largest?
- Did the size of Firefox users increase or decrease?

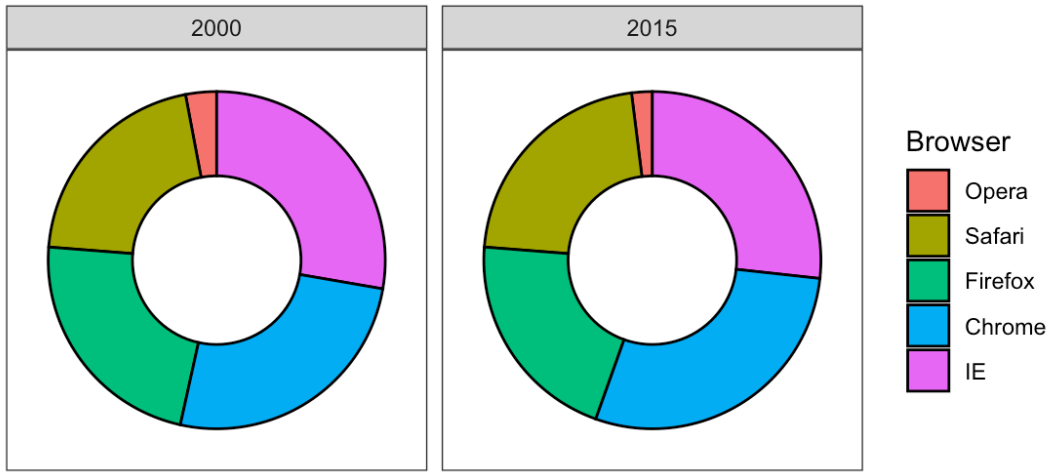
Visual Cues



Issues:

- No clear way to compare the size of each area between the two figures
- Sections are determined by both angle and area (two different dimensions)

Visual Cues



Solutions:

- Add labels to the graph
- Only use angle or area

Here is an example using a donut graph with the same data, but only using area.

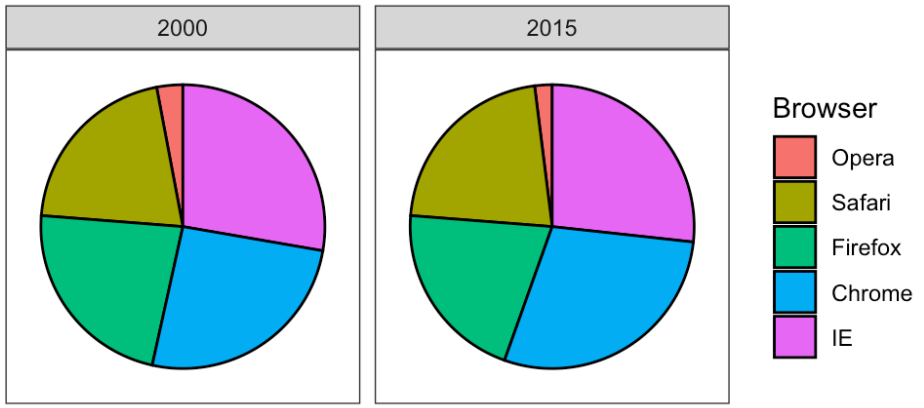
No labels -- still not clear

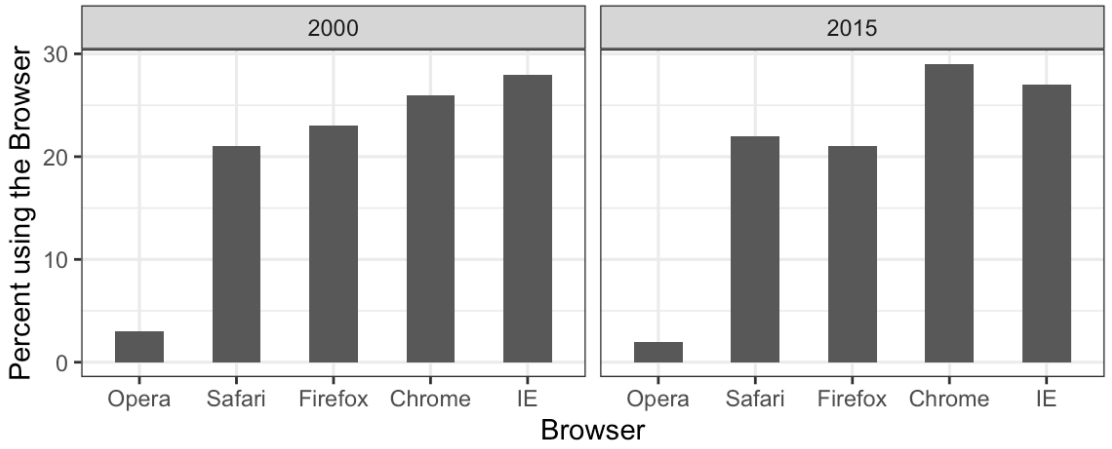
Visual Cues

Browser	2000	2015
Opera	3	2
Safari	21	22
Firefox	23	21
Chrome	26	29
IE	28	27

Sometimes, just giving the data in a table is clearer

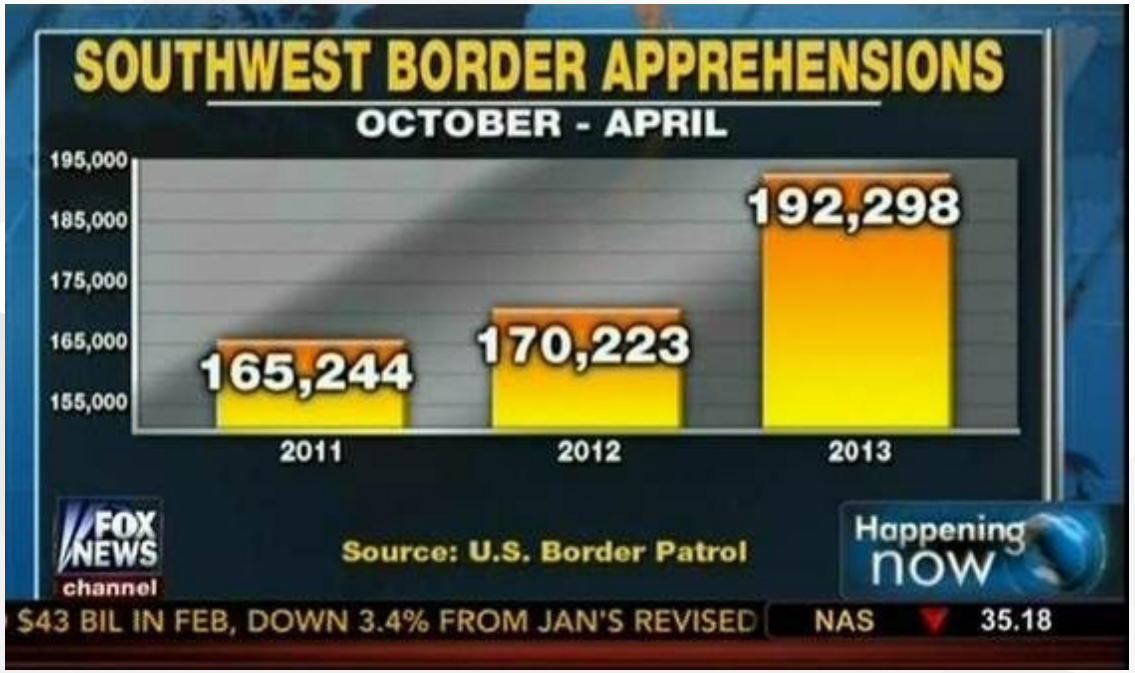
Visual Cues



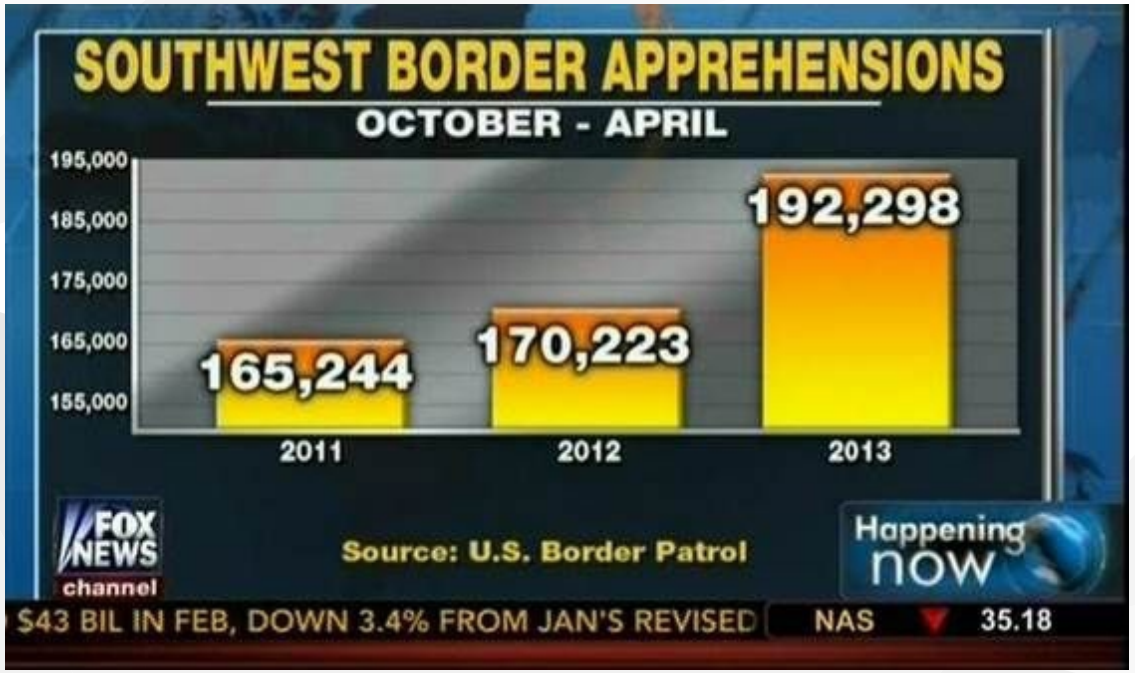


… or use a bar graph.

When to include 0



When to include 0

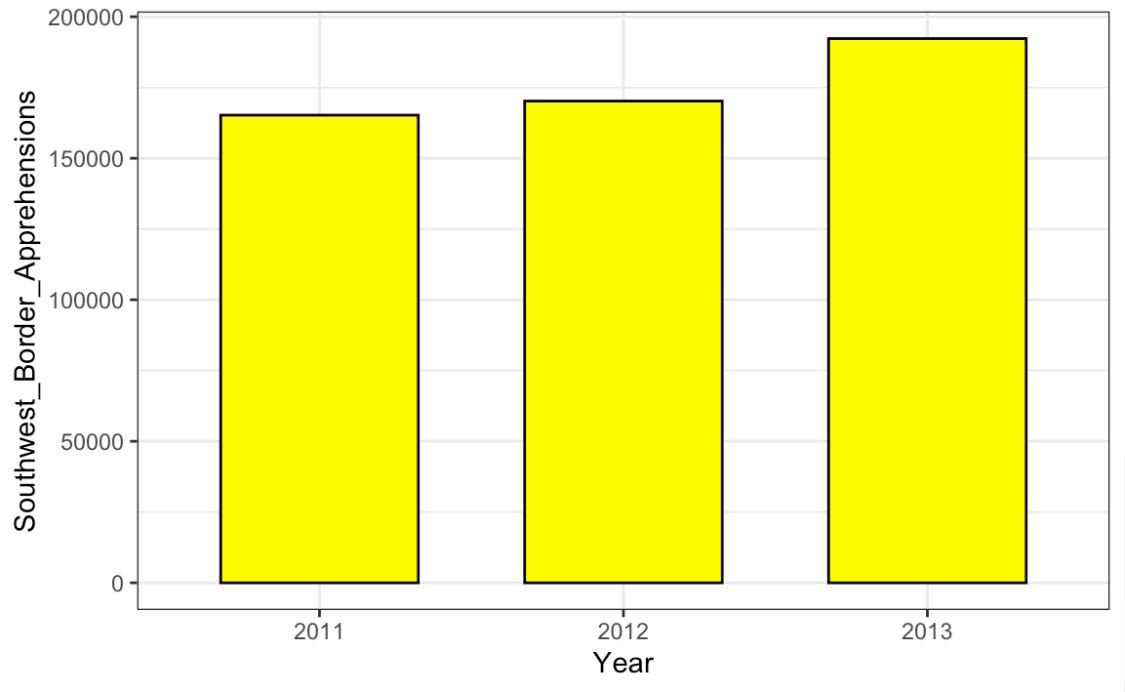


Issues:

- Lines are not proportioned correctly

It looks like 2013 has tripled 2011, but really only increased by 16%

When to include 0

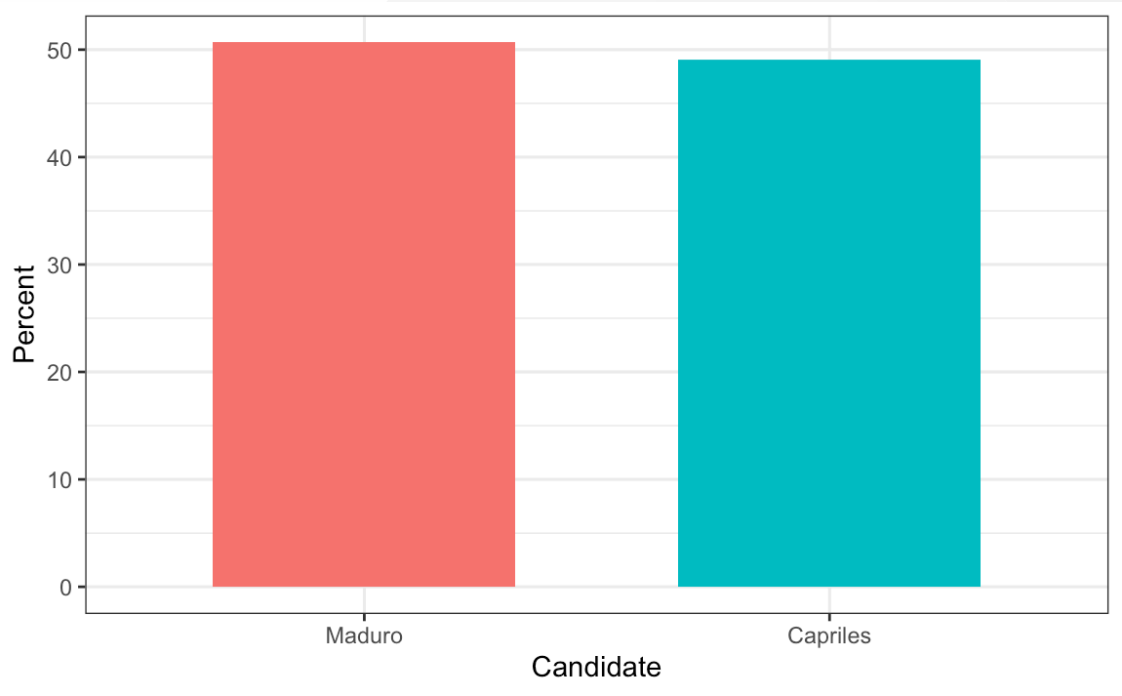


Solution:

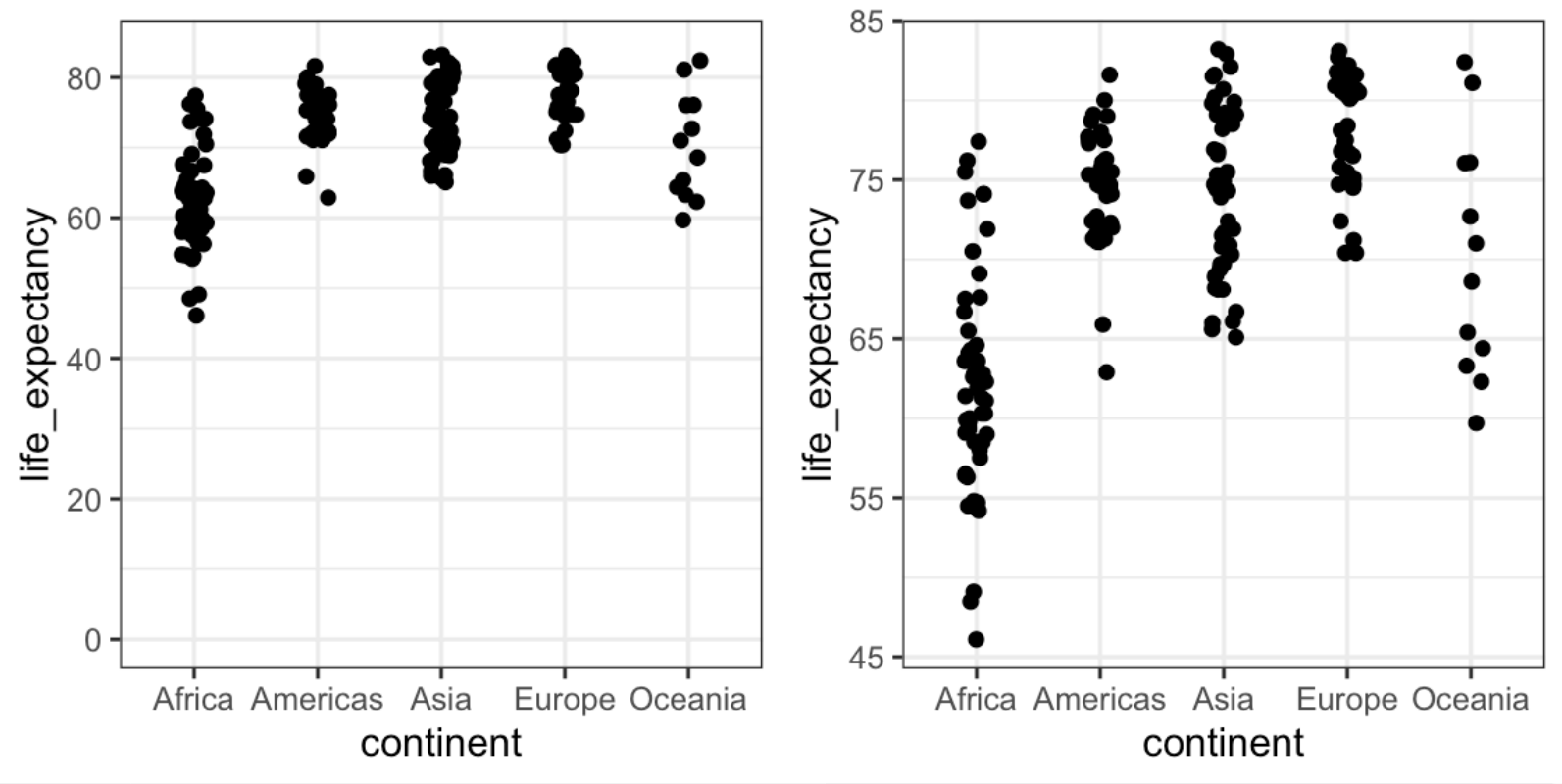
- When the distance from 0 matters, make sure 0 is displayed in the graph

When to include 0

When to include 0



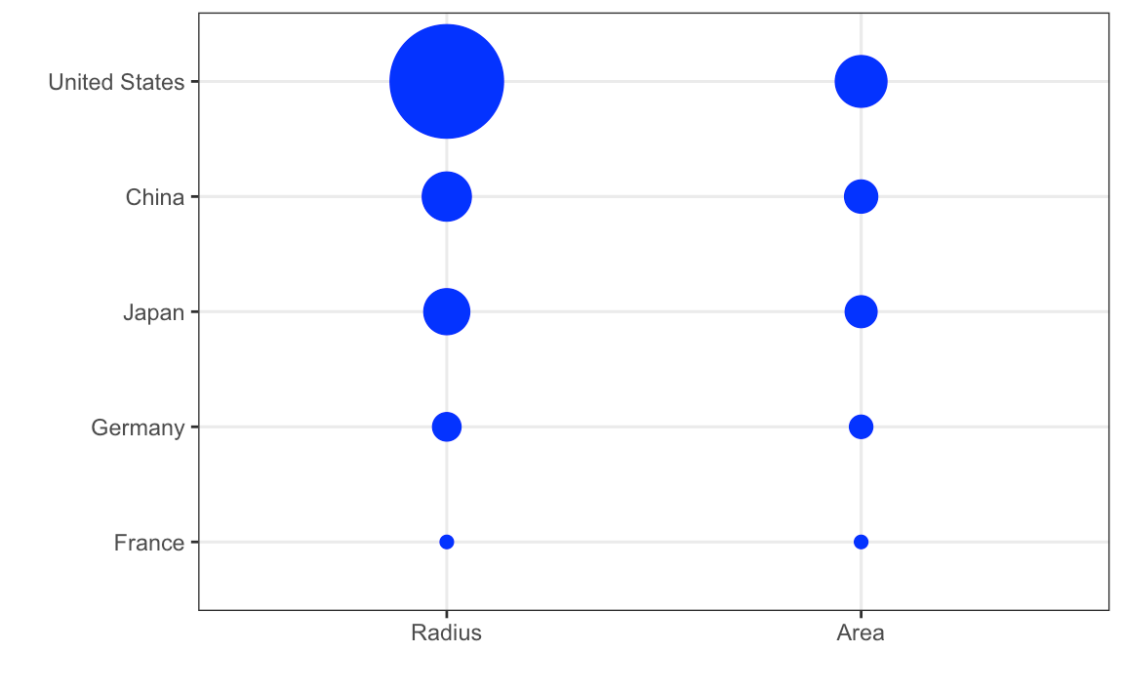
When including 0 isn't needed



In cases where we look at a distribution of values, the 0 is not really necessary

Distorting Quantities

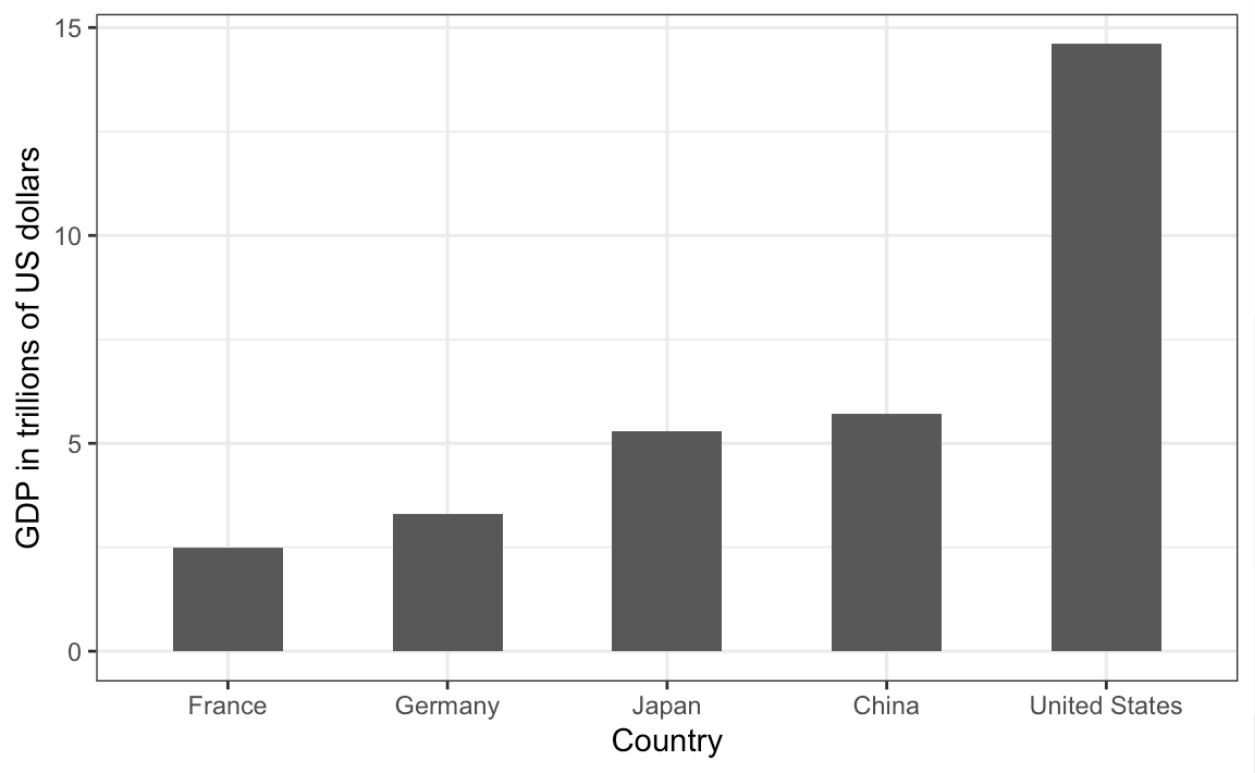
Distorting Quantities



Issues:

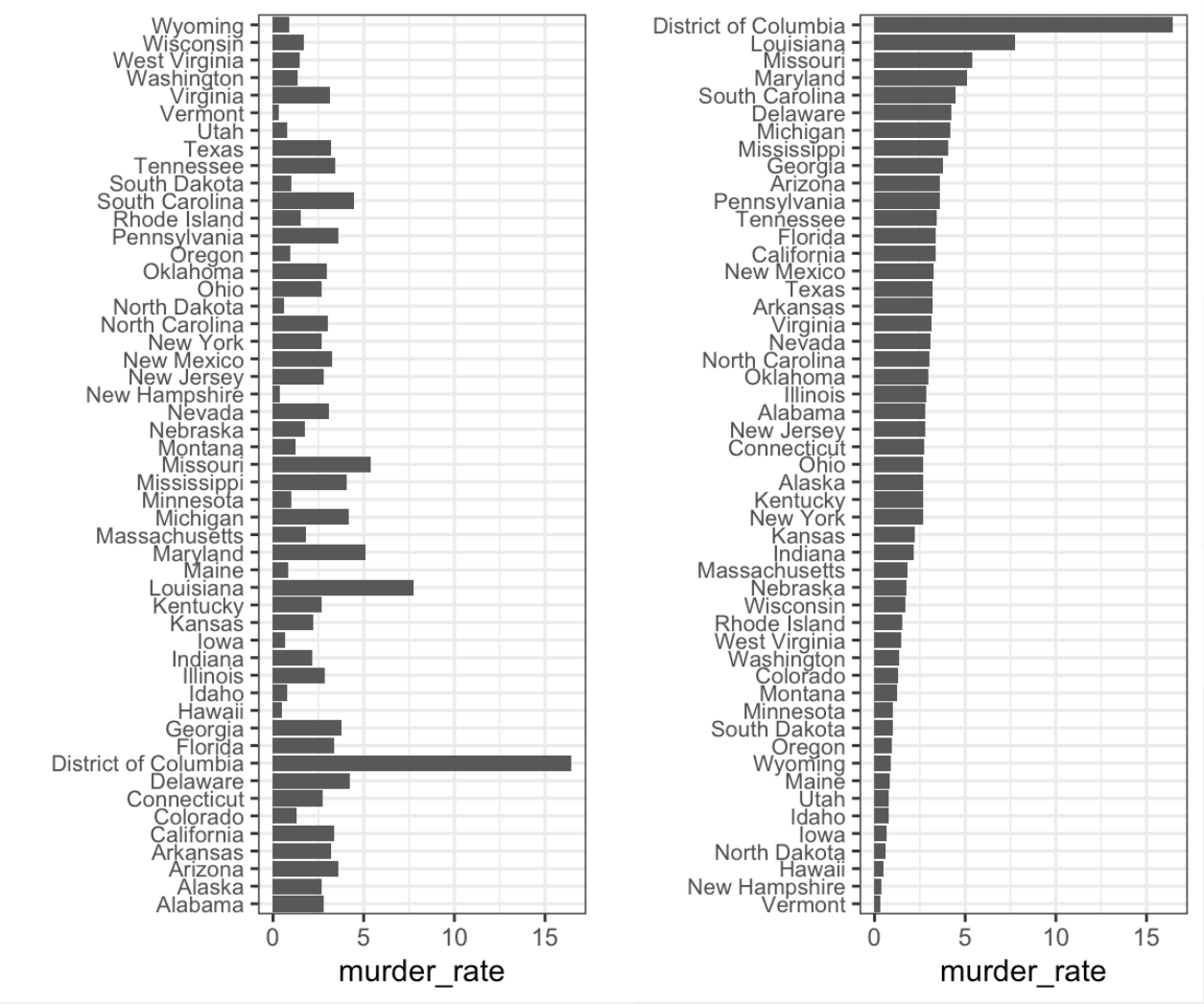
- The most obvious measure is the area (US 5x as large as China)
- Actually used radius/diameter (US 3x as large as China)

Distorting Quantities



A bar graph is easier

Meaningful Order



Meaningful Order